Best Hardware for Running OpenClaw — Mac Mini vs VPS vs Your Old PC

February 14, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For API-only OpenClaw (routing to Claude/GPT-4), almost anything works: a Raspberry Pi 5 ($65), Oracle Cloud free tier (free), or a used ThinkCentre M710Q ($85). The gateway needs 2GB RAM and one CPU core. For running local models through Ollama, you need real hardware: a Mac Mini M4 Pro with 24-48GB ($1,399-1,900) for silent low-power operation, or a used desktop with an RTX 3090 ($850+) for faster inference. The Mac Mini draws 4 watts idle ($6/year electricity). The desktop with a 3090 draws 90-130 watts idle ($142-205/year). Pick based on whether you value silence and efficiency or raw GPU speed.

📚 **More on this topic:** OpenClaw Setup Guide · OpenClaw 100% Local · Mac vs PC for Local AI · Budget Local AI PC · Build a Distributed AI Swarm

OpenClaw isn't a chatbot you open when you need it. It's an agent that runs all day. It checks your messages, fires heartbeats every 30 minutes, executes scheduled tasks, and waits for instructions on WhatsApp or Telegram. That means whatever hardware you run it on stays powered 24/7.

A gaming desktop with an RTX 3090 draws 90-130 watts at idle. That's $142-205 per year in electricity before you run a single query. A Mac Mini M4 draws 4 watts. That's $6 per year. And a free Oracle Cloud VPS costs nothing at all.

The hardware you choose determines three things: how much you spend on electricity, whether you can run local models or need cloud APIs, and how much noise you tolerate in your home. This guide covers all three paths.

Two Modes, Two Sets of Requirements

Before picking hardware, decide how you're running OpenClaw.

Mode 1: API-Only (Gateway)

The OpenClaw gateway routes your messages to a cloud LLM (Claude, GPT-4, etc.) and returns the responses. The gateway itself is a Node.js process that uses almost no resources:

Spec	Minimum	Recommended
RAM	1-2 GB	4 GB
CPU	1 core	2 cores
Storage	10 GB	20 GB
GPU	Not needed	Not needed

In this mode, a Raspberry Pi, a $4/month VPS, or a used mini PC from 2015 all work fine. The intelligence comes from Anthropic's or OpenAI's servers. You're just paying for the API calls.

Mode 2: Local Models (Gateway + Ollama)

If you want to run models locally through Ollama (no API costs, full privacy), hardware requirements jump:

Use Case	VRAM / Memory	RAM	Example Hardware
Small models (7B-8B)	6-8 GB VRAM or 16GB unified	16 GB	Mac Mini M4 16GB, PC + RTX 3060
Medium models (14B-32B)	12-24 GB VRAM or 24-48GB unified	32 GB	Mac Mini M4 Pro 24-48GB, PC + RTX 3090
Large models (70B+)	48GB+ VRAM or 64-128GB unified	64 GB+	Mac Studio M4 Max, dual GPU PC

→ Not sure what fits? Try our Planning Tool.

For the full local setup, see our OpenClaw 100% local guide. This article focuses on which physical hardware to put it on.

Path 1: Mac Mini

The Mac Mini M4 has become the default OpenClaw recommendation in the community. It's small, silent, sips power, and Apple Silicon's unified memory lets it run local models that would need a discrete GPU on a PC.

Current Pricing

Config	RAM	Storage	Price
M4 base	16 GB	256 GB	$599
M4	16 GB	512 GB	$799
M4	24 GB	512 GB	$999
M4 Pro	24 GB	512 GB	$1,399
M4 Pro	48 GB	512 GB	~$1,900

Why People Pick It

Power consumption. 4 watts idle, confirmed by independent testing. Under full LLM inference load, the M4 Pro peaks around 60-65 watts. That's less than a desktop PC draws at idle. Annual electricity cost at US average rates ($0.18/kWh): about $6.

Silence. The M4 Mini is fanless under light loads. Even under sustained AI inference, it's barely audible. If OpenClaw runs in your bedroom or living room, this matters.

Unified memory. On a PC, your GPU has 8-24GB of VRAM and models that don't fit simply don't run. On a Mac Mini with 48GB, you load a 32B model into unified memory and it runs. No spilling, no crashing. The M4 Pro's 273 GB/s memory bandwidth is slower than an RTX 3090's 936 GB/s, so token generation is slower per-byte. But loading a model that fits beats not loading one that doesn't.

macOS just works. Ollama installs in one command. No CUDA driver issues, no nvidia-smi debugging, no kernel module conflicts. MLX (Apple's framework) is even faster than Ollama for inference on Apple Silicon.

The Downsides

No NVIDIA GPU. The M4 Pro generates tokens 30-50% slower than an RTX 3090 on models that fit in 24GB VRAM. If the model fits on both, the PC with a 3090 wins on speed.

Expensive to configure. 16GB is tight for local models beyond 7B. The jump to 24GB costs $400 (M4 $999) or $800 (M4 Pro $1,399). Getting 48GB costs $1,900+. On a PC, you spend $850 on a used 3090 and get 24GB of faster VRAM.

Not upgradeable. Memory is soldered. Whatever you buy today is what you have forever. Buy too little and you're stuck. Buy too much and you overpaid.

Best Mac Mini Config for OpenClaw

API-only: M4 base 16GB ($599). More than enough. You're just running the gateway.

Local models: M4 Pro 24GB ($1,399) minimum. The M4 Pro's higher memory bandwidth (273 GB/s vs M4's 120 GB/s) makes a real difference for inference speed. 24GB loads most 14B models and some 32B models at aggressive quantization.

Path 2: VPS / Cloud Server

If you don't want hardware in your home, a VPS runs the OpenClaw gateway in a datacenter. The tradeoff: no GPU, so you're locked into API-only mode (unless you rent GPU instances, which cost $0.50-3+/hour and defeat the purpose).

Free and Cheap Options

Provider	Plan	Specs	Monthly Cost
Oracle Cloud	Always Free ARM	4 cores, 24GB RAM, 200GB storage	Free
Hetzner	CAX11	2 vCPU ARM, 4GB RAM, 40GB storage	$4.10
Hetzner	CAX21	4 vCPU ARM, 8GB RAM, 80GB storage	$7.00
AWS Lightsail	Cheapest	1 vCPU, 0.5GB RAM, 20GB SSD	$3.50

Oracle Cloud Free Tier

Oracle's always-free ARM instances are absurdly generous: 4 Ampere A1 cores, 24GB RAM, 200GB block storage. That's more than enough for OpenClaw's gateway. The catch: provisioning is difficult. High demand means instances in popular regions are frequently unavailable. You may need to try repeatedly over days or use a script that polls for availability.

If you get one, it's the best deal in this entire article. Free is free.

Hetzner

Hetzner's ARM instances are the reliable paid option. The CAX11 at $4.10/month gives you 2 ARM cores, 4GB RAM, and 40GB storage. That's plenty for the OpenClaw gateway. Datacenters in Germany and Finland only, so latency from the US is 80-120ms. For an agent that responds over WhatsApp, that's not noticeable.

Why VPS Works (And Why It Doesn't)

Works for: API-only OpenClaw where you want the gateway accessible from anywhere, don't want hardware at home, and are already paying for cloud LLM APIs anyway.

Doesn't work for: Running local models. VPS instances don't have GPUs. GPU cloud instances (Lambda, RunPod, etc.) cost $0.50-3+ per hour. Running one 24/7 costs $360-2,160/month. At that point, buy a Mac Mini.

Doesn't work for: Privacy-focused users. Your messages route through a third-party datacenter. The LLM calls go to another third party. Nothing stays on your hardware.

Path 3: Repurposed PC

This is the path for people who already have hardware sitting around, or who want flexibility on a budget.

The Mini PC Route (API-Only)

Used enterprise mini PCs are ridiculously cheap because businesses lease them, use them for 3-4 years, and dump them by the pallet:

Machine	Typical eBay Price	CPU	Idle Power
Lenovo ThinkCentre M710Q	$70-100	i5-7500T	11-14W
Dell Optiplex 3060 Micro	$80-120	i5-8500T	11-14W
HP EliteDesk 800 G4 Mini	$100-150	i5-8500T	12-18W

These machines are about the size of a paperback book. Silent under light loads. 11-14 watts at idle. Ubuntu installs cleanly. Perfect for running the OpenClaw gateway and routing to cloud APIs.

The ThinkCentre M710Q at $85 is what I use as a light-inference node in my distributed swarm build. It runs embeddings and small model queries on 35 watts. For API-only OpenClaw, it barely breaks a sweat.

The Desktop Route (Local Models)

If you want to run local models through Ollama, you need a GPU. That means a desktop (or at least a tower-style case) with a PCIe slot:

Setup	Cost	What You Get
Used Optiplex tower + RTX 3060 12GB	$250-400	7B-13B models, 12GB VRAM
Used workstation + RTX 3090	$850-1,200	32B models, 24GB VRAM
Existing gaming PC + 3090	$850 (GPU only)	32B models, 24GB VRAM

The budget PC build guide covers this path in detail. The short version: a used Dell Optiplex tower ($100-150) plus a used RTX 3060 12GB ($170-200) gets you a working local AI machine for under $450.

The Downside: Power and Noise

A desktop with an RTX 3090 draws 90-130 watts at idle. Under inference load, 400-450 watts. That's a space heater. Annual idle electricity at US average rates: $142-205. And the GPU fans spin, especially under load. If this machine lives next to your desk, you'll hear it.

Compare that to a Mac Mini at $6/year or a ThinkCentre at $20/year. The desktop's electricity bill can exceed the cost of the hardware itself within two years.

Raspberry Pi 5 (The $65 Gateway)

The Pi 5 works as an OpenClaw gateway. Quad-core ARM Cortex-A76 at 2.4GHz, available with 2-8GB RAM. The 4GB model ($85) is the sweet spot for OpenClaw's Node.js process.

Pi 5 Model	RAM	Price
Pi 5 2GB	2 GB	$65
Pi 5 4GB	4 GB	$85
Pi 5 8GB	8 GB	$125

Power draw: 3-5 watts. Annual electricity: $5-8. Use the official 5V/5A USB-C power supply for 24/7 reliability, and boot from an NVMe SSD (via the Pi's PCIe slot) instead of a microSD card. SD cards degrade under constant writes.

The Pi won't run local models. It's gateway-only. But at $85 total cost and $7/year electricity, it's the cheapest physical hardware option that works.

Power Consumption: The Hidden Cost

OpenClaw runs 24/7. Electricity costs add up. Here's what each option actually draws and costs over a year at the US average of $0.18/kWh:

Hardware	Idle Power	Annual Electricity	3-Year Electricity
Mac Mini M4	4W	$6	$19
Raspberry Pi 5	3-5W	$5-8	$15-24
ThinkCentre M710Q	11-14W	$17-22	$52-66
Dell Optiplex Micro	10-15W	$16-24	$47-71
Desktop + RTX 3090	90-130W	$142-205	$426-616
VPS (Oracle free)	0W (your end)	$0	$0
VPS (Hetzner CAX11)	0W (your end)	$49 (hosting)	$148 (hosting)

The Mac Mini and Pi 5 are in a different category. At 4 watts, the Mac Mini costs less to run for a year than a single dinner out.

The desktop with a 3090 is the outlier. If you run it as an always-on OpenClaw server, electricity alone costs more than a Mac Mini M4 base within two years. The 3090 makes sense for active inference workloads (running models when you need them), not as a 24/7 idle server.

Recommendations by Budget

$0: Use What You Have

Option A: Oracle Cloud free tier. 4 ARM cores, 24GB RAM, free forever. Run the OpenClaw gateway, connect to Claude or GPT-4 via API. The hardest part is provisioning the instance (high demand, limited availability). Once you get one, it's the best deal here.

Option B: Old laptop or desktop. Any machine from the last decade with 4GB+ RAM and a wired Ethernet connection works as an API-only gateway. Install Ubuntu Server, run OpenClaw, leave it on. The electricity cost depends on the machine, but even a 40W laptop is only $63/year.

$50-100: Used Mini PC

A ThinkCentre M710Q ($85 on eBay) or Dell Optiplex Micro ($80-120) running Ubuntu Server. API-only. 11-14 watts idle. Small, quiet, reliable. This is the practical choice for most API-only users who want physical hardware.

Add a monitor-less headless setup (SSH in from your main machine) and it disappears into a shelf or behind a router.

$500: Two Good Options

Option A: Mac Mini M4 base ($599). API-only, but future-proofed with 16GB unified memory. Can run 7B models locally through Ollama if you want to experiment, though 16GB is tight for anything larger. Silent, 4W idle, macOS.

Option B: Budget PC + RTX 3060 12GB (~$400). A used Optiplex tower plus a used 3060 gives you 12GB VRAM for local models up to 13B. Louder and power-hungry compared to the Mac Mini, but you get real GPU inference. See the budget build guide for the full parts list.

$1,000+: Local Models, Always-On

Option A: Mac Mini M4 Pro 24GB ($1,399). The sweet spot for local OpenClaw. 24GB unified memory loads 14B models comfortably and some 32B models at Q4. 273 GB/s memory bandwidth. 4W idle. Silent. This is what most OpenClaw power users on Mac are running.

Option B: Used desktop + RTX 3090 (~$850-1,200). 24GB VRAM, faster inference than the Mac on models that fit. Runs 32B models at Q4 with room for context. Downside: 90-130W idle, fan noise, needs a case and PSU that fit a 350W three-slot GPU. Best if you already own a compatible desktop and just need the GPU.

The tradeoff: The Mac Mini is quieter, more efficient, and easier to set up. The 3090 is faster at inference and has more headroom for large context windows. If it runs in your living space, get the Mac. If it runs in a closet or garage, the 3090 is more capable per dollar.

When to Run Local Models vs Just Use APIs

This is the decision that determines your hardware requirements more than anything else.

Factor	Run Local	Use APIs
Monthly budget	$0 (already own hardware)	$15-150/month (API costs)
Privacy	Data never leaves your machine	Messages route through cloud providers
Model quality	Good (32B local) to great (70B on Mac)	Best available (Claude Opus, GPT-4)
Inference speed	15-45 tok/s (depends on hardware)	50-100+ tok/s
Reliability	Depends on your hardware and power	99.9%+ uptime
Upfront cost	$600-1,400	$0-85 (gateway hardware)
Complexity	Install Ollama, manage models, tune configs	Point at API, done

The Hybrid Approach

Most practical OpenClaw setups use both. The token optimization guide covers this in detail, but the short version:

- **Heartbeats** (every 30 minutes, just keep-alive pings): route to a free local model via Ollama
- **Simple tasks** (file management, reminders, basic queries): route to a cheap cloud model (Haiku at $1/$5 per million tokens)
- **Complex tasks** (multi-step reasoning, code generation, research): route to Claude Sonnet or Opus

This lets you run the gateway on cheap hardware (mini PC, Pi, VPS) with Ollama handling only heartbeats. Complex work goes to the cloud. You get the cost savings of local heartbeats without needing a $1,400 Mac or an 3090 to run the smart model.

See the model routing guide for the exact `openclaw.json` configuration.

Bottom Line

The hardware decision comes down to one question: are you running local models or not?

API-only (no local models): Get the cheapest thing that stays on. Oracle free VPS, a $65 Raspberry Pi 5, or an $85 used ThinkCentre. The gateway needs almost nothing. Spend your budget on API credits instead of hardware.

Local models (privacy, no recurring costs): Get a Mac Mini M4 Pro 24GB ($1,399) for silent, efficient, always-on operation. Or get a used desktop with an RTX 3090 ($850+) for faster inference at the cost of noise and power draw.

Not sure yet: Start API-only on cheap hardware. If you decide you want local models later, you can always add them. The gateway hardware doesn't change. You just add Ollama and a capable machine to run it on.

Related Guides

- OpenClaw Setup Guide — installation from scratch
- Running OpenClaw 100% Local — zero API costs with Ollama
- OpenClaw Token Optimization — hybrid routing for 97% cost reduction
- OpenClaw Model Routing — route tasks to the right model
- Mac vs PC for Local AI — deep dive on the platform question
- Build a Local AI PC for Under $500 — the desktop route
- Building a Distributed AI Swarm — multi-node setup
- How Much VRAM Do You Need? — model sizing by hardware

Get notified when we publish new guides.

Subscribe — free, no spam
